LLM3D: a log-linear modeling-based method to predict functional gene regulatory interactions from genome-wide expression data

Geeven, G.; MacGillavry, H.D.; Eggers, R.; Sassen, M.M.; Verhaagen, J.; Smit, A.B.; de Gunst, M.C.M.; van Kesteren, R.E.

published in
Nucleic Acids Research
2011

DOI (link to publisher)
10.1093/nar/gkr139

document version
Publisher's PDF, also known as Version of record

Published online 21 March 2011

Nucleic Acids Research, 2011, Vol. 39, No. 13 ***5313–5327***
doi:10.1093/nar/gkr139

LLM3D: a log-linear modeling-based method to predict functional gene regulatory interactions from genome-wide expression data

Geert Geeven[1], Harold D. MacGillavry[2], Ruben Eggers[3], Marion M. Sassen[2], Joost Verhaagen[1,3], August B. Smit[2], Mathisca C. M. de Gunst[1] and Ronald E. van Kesteren[2,*]

[1]Department of Mathematics, Faculty of Sciences, VU University, De Boelelaan 1081, 1081 HV Amsterdam, [2]Department of Molecular and Cellular Neurobiology, Center for Neurogenomics and Cognitive Research, Neuroscience Campus Amsterdam, VU University, De Boelelaan 1085, 1081 HV Amsterdam and [3]Department of Neuroregeneration, Netherlands Institute for Neuroscience, Meibergdreef 47, 1105 BA Amsterdam, The Netherlands

Received October 25, 2010; Revised February 24, 2011; Accepted February 25, 2011

ABSTRACT

All cellular processes are regulated by condition-specific and time-dependent interactions between transcription factors and their target genes. While in simple organisms, e.g. bacteria and yeast, a large amount of experimental data is available to support functional transcription regulatory interactions, in mammalian systems reconstruction of gene regulatory networks still heavily depends on the accurate prediction of transcription factor binding sites. Here, we present a new method, log-linear modeling of 3D contingency tables (LLM3D), to predict functional transcription factor binding sites. LLM3D combines gene expression data, gene ontology annotation and computationally predicted transcription factor binding sites in a single statistical analysis, and offers a methodological improvement over existing enrichment-based methods. We show that LLM3D successfully identifies novel transcriptional regulators of the yeast metabolic cycle, and correctly predicts key regulators of mouse embryonic stem cell self-renewal more accurately than existing enrichment-based methods. Moreover, in a clinically relevant *in vivo* injury model of mammalian neurons, LLM3D identified peroxisome proliferator-activated receptor γ (PPARγ) as a neuron-intrinsic transcriptional regulator of regenerative axon growth. In conclusion, LLM3D provides a significant improvement over existing methods in predicting functional transcription regulatory interactions in the absence of experimental transcription factor binding data.

INTRODUCTION

Insight into gene regulatory networks is crucial for the understanding of biological systems under normal and pathological conditions. An important step in the analysis of gene networks is the prediction of functional transcription factor binding sites (TFBSs) within gene regulatory sequences. Recently, advanced methods have been developed to predict TFBSs *in silico* (1–7). Public databases containing large collections of experimentally validated binding sites can be used to derive probabilistic models of TFBSs and software algorithms can subsequently be employed to scan potential gene regulatory sequences for the prediction of new sites. However, in contrast to simple model organisms such as yeast, mammalian gene regulatory sequences are often large and can be located up to several thousands of base pairs away from transcription start sites. Consequently, mammalian

*To whom correspondence should be addressed. Tel: +31 20 5987111; Fax: +31 20 5989281; Email: ronald.van.kesteren@cncr.vu.nl
Present address:
Harold D. MacGillavry, Department of Physiology, University of Maryland School of Medicine, Baltimore MD, USA.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

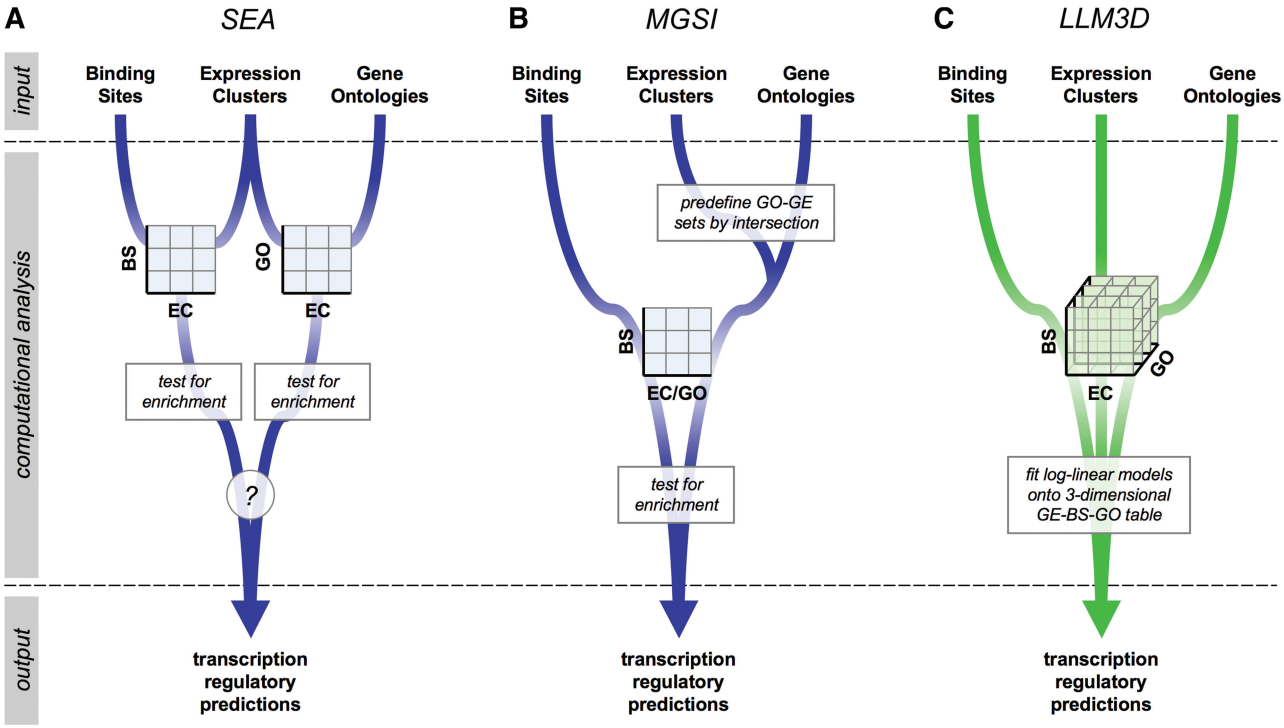


Figure 1. Comparison of LLM3D with other gene set enrichment analysis approaches. (**A**) In singular enrichment analysis (SEA), gene expression clusters (EC) are independently tested for enrichment of binding sites (BS) and gene ontology (GO) terms using two 2D contingency tables. It is not clear how meaningful relationships between the two should be inferred. (**B**) In multigene set intersection (MGSI), multiple gene sets are predefined based on intersecting sets of co-expressed genes with sets of genes sharing GO terms. MGSI considers all three variables in a single 2D contingency table in which gene expression and GO data are collapsed into a single combined variable. (**C**) In LLM3D, gene expression, binding site data and GO annotations are used as separate input variables in a single 3D contingency table. To this table, log-linear models are fitted in order to test the joint dependence of all three variables simultaneously.

TFBS predictions are usually less accurate and more likely to contain false positives. A reduction in false positive TFBS predictions can be achieved by improving the quality of the biological input data, for instance by considering TF binding affinities (8,9), TF cooperativity at *cis*-regulatory modules (10,11) or evolutionary conservation of binding sites across species (12,13), or by improving the way in which computational methods make use of these data.

A common method to reduce false positive TFBS predictions at the computational level involves the identification of TFBSs that are enriched in subsets of related genes compared to a control (background) set of genes (14,15). Co-regulation and co-functionality are often used as criteria to define gene sets of interest. In order to study enrichment of both TFBSs and gene function in co-expressed genes, two different computational approaches can be used. The first approach, referred to as singular enrichment analysis (SEA) (14), allows separate quantification of gene ontology (GO) term and TFBS enrichment in sets of co-expressed genes (Figure 1A). SEA typically returns separate lists of enriched GO terms and TFBSs (16,17), but is not designed to predict transcriptional targets using gene expression, TFBS and GO data simultaneously. The second approach, which we will refer to as the multigene set by intersection (MGSI) approach, predefines multiple sets of co-expressed genes sharing the same GO term, and subsequently tests each set for TFBS enrichment (Figure 1B). MGSI-based methods provide a significant improvement over SEA and perform better in predicting functional TFBSs (18,19). However, MGSI collapses gene expression and GO annotation into a single combined variable. As a result, important information about the joint dependence of all three variables (i.e. gene expression, GO association and TFBS presence) is lost.

We present a novel method that uses log-linear modeling of 3D contingency tables (LLM3D), to identify experiment-specific associations between gene expression, TFBS presence and gene function (Figure 1C). We show that LLM3D provides a significant improvement over existing methods. We validate our method using published genome-wide gene expression and transcription factor binding data, and demonstrate that the gene regulatory predictions made by LLM3D have significantly higher predictive value compared with MGSI, and are biologically relevant, both in yeast and in mammals. Finally, we showcase LLM3D by performing and analyzing a genome-wide expression profiling study in a clinically relevant animal model for the functional regeneration of injured neurons. *Post hoc* experimental validation shows that in this case LLM3D is able to identify functional gene regulatory interactions that remain undetected using existing methodologies.

MATERIALS AND METHODS

LLM3D

Here, we give a brief outline of LLM3D; a detailed description can be found in the Supplementary Methods. For each TFBS–GO pair of interest, LLM3D cross-classifies all genes according to observed gene expression, GO annotation and TFBS prediction to obtain a 3D table (see Fig. 2B for an example). The rows of this table correspond to the GO terms, the columns to the TFBSs, and the gene expression clusters define the layers of the table. Let μ_{ijk} denote the expected number of genes in row i, column j and layer k. Then, for a sample of genes of size N and under the null hypothesis of complete independence between rows, columns and layers:

$$\log \mu_{ijk} = \eta + \alpha_i + \beta_j + \gamma_k.$$

This model is called the null model ($M^{(0)}$). Under the assumption that the null model holds, and under multinomial sampling, the likelihood of the observed data is completely determined by the unknown parameters, which can be estimated using maximum likelihood. Lack of fit can be formally tested using a standard likelihood ratio G^2 statistic (20). For a 3D contingency table, there are eight other natural models to consider. These models differ in the parameters used to describe the expected counts and the dependence relationships they imply between the rows, columns and layers of the table (see Supplementary Methods for details). For each of these models, we estimate the parameters using maximum likelihood and calculate the G^2 statistic. Next, we select the model that best describes the observed data using Akaike's information criterion (AIC) (21), which can be calculated from G^2 and the degrees of freedom of the model. For re-analysis of yeast metabolic cycle data and mouse ES cell data, we considered all models with at least two two-way (first order) interactions, i.e. $M^{(4)}$, $M^{(5)}$, $M^{(6)}$, $M^{(7)}$ and $M^{(S)}$. For analysis of the neuronal regeneration data we only considered models with all pairwise interactions and the saturated model, i.e. $M^{(7)}$ and $M^{(S)}$.

Selection and visualization of biologically relevant TFBSs

An enrichment score is used to quantify the relative enrichment of target genes in different expression clusters and to filter and visualize LLM3D results. For K different expression clusters, the enrichment of target genes that belong to a certain GO class and have a certain TFBS is calculated as follows. For $k = 1,\ldots K$, let n_k denote the observed number of genes in the corresponding cell of the table, and $m_k{}^{M(0)}$ the expected number of genes in that cell under the assumption that model $M^{(0)}$ holds. We then use

$$e_k = \frac{n_k - \hat{m}_k^{M(0)}}{\sqrt{\hat{m}_k^{M(0)}}}$$

as a measure of enrichment of target genes in cluster k for a TFBS–GO pair of interest. Values of e_k with a positive sign indicate enrichment, whereas a negative sign indicates depletion. The set of predicted target genes for a given TFBS–GO pair is then defined as the union of sets of TFBS–GO genes in all clusters with a positive e_k. For any two clusters k_1 and k_2 of interest, relative enrichment is assessed in cluster k_1 with respect to k_2 using a score s that compares e_{k1} and e_{k2}, where

$$s_{k_1k_2} = \begin{cases} 0, & \text{if } e_{k_1} < 0 \text{ and } e_{k_2} > 0 \\ 0.5, & \text{if } e_{k_1} < 0 \text{ and } e_{k_2} < 0 \\ e_{k_1}/(e_{k_1}+e_{k_2}), & \text{if } e_{k_1} > 0 \text{ and } e_{k_2} > 0 \\ 1, & \text{if } e_{k_1} > 0 \text{ and } e_{k_2} < 0 \end{cases}$$

Next, all significant TFBSs predicted by LLM3D are ranked according to the sample variance of their $s_{k_1k_2}$ scores over all associated GO terms, and the top-ranked TFBSs are then selected as the ones with the most between-cluster-specific regulatory potential. The $s_{k_1k_2}$ scores can be visualized in a heat map (see for example Figure 5B).

MGSI

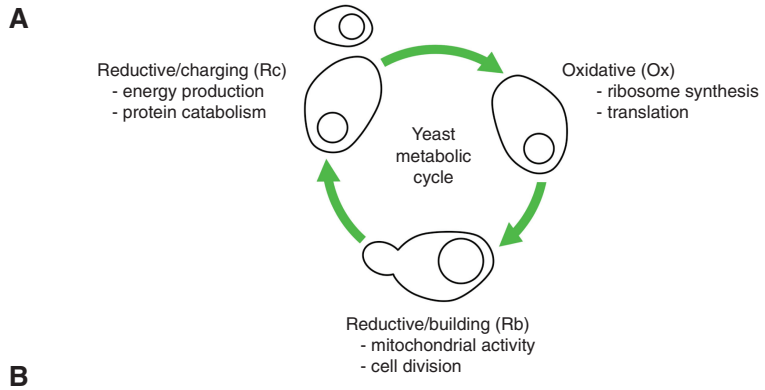
For any given gene expression cluster and GO term, MGSI first generates a new gene set by intersecting the genes in the expression cluster with the set of genes annotated to the GO term. Enrichment of any TFBS in this new set is tested using a Fisher's exact test (one-sided) for 2D contingency tables. A Benjamini Hochberg correction is applied to the resulting P-values to correct for multiple testing with the aim of controlling the false discovery rate (FDR) at 10%.

Yeast TFBS annotation

Yeast ORF sequences with introns and untranslated regions 1000 bp immediately upstream of the initial ATG were downloaded from the Saccharomyces Genome Database (SGD) on http://www.yeastgenome.org. Log-odds matrices representing probabilistic models for binding sites, 124 in total, were downloaded from http://fraenkel.mit.edu/improved_map/ and converted to probability matrices to be used with the Motifscanner tool (1). Motifscanner was used to computationally predict binding sites for all TFs on both DNA strands with the 'prior probability' parameter set to 0.15. A third-order Markov background model was generated, trained on the SGD sequences with the accompanying CreateBackgroundModel tool.

Mammalian TFBS annotation

Gene regulatory sequences (5000 bp upstream to 2000 bp downstream of the predicted transcription start site) for all human, mouse and rat genes identifiable by Entrez Gene ID were downloaded using the biomaRt package under R. Potential TFBSs were predicted *in silico* using all 214 vertebrate non-redundant position weight matrices in the TRANSFAC Professional database (release 11.1) (22) and the supplied MATCH-tool (5) with parameters set to minimize false positives. The MATCH output was used to create a binary matrix with rows corresponding to regulatory sequences and columns corresponding to TRANSFAC matrices. In this matrix, '1' represents



B

MGSI

	ACE2 P	ACE2 A
NOT [Ox AND mitosis]	596	4880
[Ox AND mitosis]	1	22

$p = 0.27$

	ACE2 P	ACE2 A
NOT [Rb AND mitosis]	589	4853
[Rb AND mitosis]	8	49

$p = 0.27$

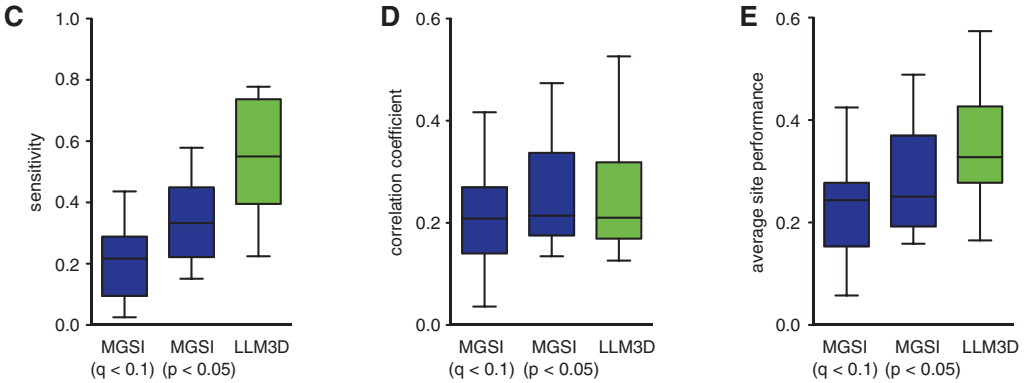
	ACE2 P	ACE2 A
NOT [Rc AND mitosis]	592	4872
[Rc AND mitosis]	5	30

$p = 0.33$

LLM3D

	ACE2 Not periodic P	Not periodic A	Ox P	Ox A	Rb P	Rb A	Rc P	Rc A
NOT mitosis	231	1945	87	875	106	764	149	1158
mitosis	10	59	1	22	8	49	5	30

$p = 0.00002$



F

Cluster	TF	Associated GO-terms
Ox	ABF1 CBF1 RAP1 FHL1	ribosome biogenesis, rRNA metabolic process, RNA processing, regulation of translation
Rb	ABF1 HAP1 CBF1 SKN7	mitochondrion organization and biogenesis, energy derivation by oxidation of organic compounds, oxidation reduction, phosphate metabolic process, post-translational protein modification
Rc	CBF1 REB1 FKH1 UME6	response to chemical stimulus, cellular protein catabolic process, proteolysis, post-translational protein modification, protein ubiquitination

Figure 2. LLM3D correctly infers gene regulatory interactions from yeast metabolic cycle gene expression data. (**A**) Schematic representation of the yeast metabolic cycle. The three phases of the cycle (Ox, Rb and Rc) are indicated together with the biological processes that dominate each phase. (**B**) Example contingency tables demonstrating the LLM3D approach. LLM3D detects a significant interaction of ACE2 binding sites with 'mitosis' GO genes in yeast metabolic cycle gene expression clusters, whereas MGSI does not (P: present; A: absent; see text for details). The highest enrichment of ACE2/'mitosis' genes is observed in the Rb expression cluster (green box), which corresponds with the mitotic phase of the cycle. (**C**) Sensitivity (Sn) of LLM3D and MGSI with respect to predicting transcriptional regulators in the yeast metabolic cycle. LLM3D shows higher Sn values compared with MGSI, even when the stringency of the latter is reduced to a *P*-value cut-off of 0.05 without correction for multiple testing. (**D**) Correlation coefficient (CC) of LLM3D and MGSI predictive performances. (**E**) Average site performance (ASP) of LLM3D and MGSI. (**F**) List of the top-four LLM3D-predicted TFs per expression cluster for which at least five additional true positive targets are predicted compared with MGSI, together with their associated GO-terms.

the presence of at least one predicted TFBS, whereas '0' represents the absence of predicted TFBSs. In addition, all human, mouse and rat genes in LLM3D were also annotated with human/mouse/rat (HMR) conserved TFBSs downloaded from http://genome.ucsc.edu/. This allows LLM3D analysis to be limited to evolutionary conserved binding sites only. For the mouse ES cell data analysis, 12 additional TFBS motif models were used that were derived from chromatin immunoprecipitation (ChIP)-Seq data (23). FIMO, part of the MEME software suite (24), was used to predict TFBS occurrences for these 12 TFs in regulatory sequences of all known mouse RefSeq genes (UCSC, NCBI36/mm8 assembly).

GO pre-selection

Yeast GO annotation data were extracted from the R-package 'org.Sc.sgd.db', which was downloaded from http://www.bioconductor.org. GO biological process annotations for human, mouse and rat genes were retrieved from http://www.geneontology.org/. Informative GO terms were selected as follows. For any GO term *i,* let GO(i) be the set of genes whose annotation contains term i and let $N(i)$ be the size of that set. We let Child(i) denote the set of children of i in the directed acyclic GO graph. Let $M(i)$ be the maximum over $N(r)$, for terms r in Child(i). For any positive number γ and any term i, we now say that i is the most informative GO term at level γ if $N(i) \geq \gamma$ and $M(i) < \gamma$. For analysis of the yeast metabolic cycle data, we considered all most informative GO terms at level 20. For the analysis of the mouse ES cell data and the rat neuronal regeneration data, we selected most informative GO terms at level 50. As an alternative to reduce redundancy in the selection of GO terms, the LLM3D R-package allows for the use of GO slim terms.

Yeast metabolic cycle data

For analysis of the yeast metabolic cycle data, we used the original clustering (25). The MRM refined regulatory map providing true interactions between TFs and target genes based on ChIP-chip data (26) was downloaded from http://fraenkel.mit.edu/improved_map/. True TF–target gene interactions reported in the YEASTRACT database (27) were downloaded from http://www.yeastract.com. For validation of predicted regulatory interactions we used a 'RegulationMatrix' containing all documented regulatory interactions in eiter MRM or YEASTRACT.

Mouse embryonic stem cell data

For analysis of the mouse embryonic stem (ES) cell data, we used the gene expression clusters defined by Ouyang et al. (28). TF association strength (TFAS) scores as computed by Ouyang et al. (28) were used for all RefSeq genes to define target genes for all 12 TFs. Genes with a positive TFAS were defined as true targets.

Animals and surgical procedures

Adult Wistar rats (~220 g; Harlan, The Netherlands) were subjected to either sciatic nerve (SN) or dorsal root (DR) crush as described previously (29) and in approval with the KNAW animal experimentation committee for animal welfare. L4–6 DRGs were isolated at 12, 24, 72 h and 7 days after surgery. Animals were independent from our previous microarray study, and we used three animals per time point in order to obtain three independent biological replicates. Control DRGs were obtained from three uninjured animals.

Microarray hybridization, normalization and analysis

Total RNA was isolated from L4, L5 and L6 DRGs using Trizol (Invitrogen; Carlsbad, CA, USA). RNA from three control animals was pooled prior to the labeling to serve as a common reference sample. RNA samples were amplified, labeled and hybridized to Agilent 44K rat whole-genome expression arrays using standard Agilent protocols (Agilent; Santa Clara, CA, USA). Arrays were scanned using an Agilent scanner and data were read using Agilent Feature Extraction software. Array data were further processed using the R-packages Bioconductor (30) and limma (Linear Models for Microarray Data) (31) for standard background subtraction and loess normalization. For statistical analyses, we used the Bayesian approach for microarray time course data developed by Angelini *et al.* (32,33). This algorithm is implemented in a Matlab executive, termed Bayes analysis of time-series (BATS). Heat maps and hierarchical clusters were generated using TIGR MeV software (http://www.tm4.org/mev.html). Primary microarray data have been submitted to GEO (http://www.ncbi.nlm.nih.gov/geo/; accession number GSE 21007).

Expression clusters

The log-fold gene expression change (relative to control; averaged over three replicates per time point) in both experiments (SN and DR crush) was calculated for each gene. Expression data from significantly regulated genes following SN and DR crush were analyzed separately using a standard principal component analysis algorithm under R. For each gene, we used the coefficient corresponding to the first principal component to further define two homogeneous gene expression clusters: one containing genes that are either up-regulated after DR crush or down-regulated after SN crush (DR > SN), and one containing genes that are either up-regulated after SN crush or down-regulated after DR crush (SN > DR). For a small group of genes that were significantly regulated following both crushes and for which the dominant direction of log-fold change (i.e. either up- or down-regulation) coincided in both experiments, we compared the average log-fold change of expression following SN and DR crush directly for classification into (DR > SN) or (SN > DR).

Cell culture and neurite outgrowth assays

F11 cells were maintained as previously described (29). For pharmacological stimulations F11 cells were plated in 96-well plates. Medium was replaced with DMEM containing 0.5% FCS, and 1 μM of PPAR agonist (ciglitazone for PPARγ and Wy-14643 for PPARα) or antagonist (GW9662 for PPARγ and GW6471 for PPARα; all from

Sigma-Aldrich, St Louis, MO, USA) was added. Cells were fixed 2 days later and stained with anti-beta-III-tubulin (Sigma-Aldrich). Neurite outgrowth was quantified using a Cellomics ArrayScan HCS Reader and the Neuronal Profiling 3.5 Bioapplication (Cellomics Inc., Pittsburgh, PA, USA). Per well 500–1000 cells were analyzed and neurite total length per cell was calculated. For statistical analysis, the average neurite total length for five wells was compared between experimental and control conditions and a Student's *t*-test was used to determine significance. Eperiments were replicated at least four times. Dissection and dissociation of primary adult DRG neurons was carried out as described (29). After 40 h in culture neurons, were fixed and immunostained with anti-beta-III-tubulin. The longest neurite of each of 100–200 neurons was measured using the ImageJ Simple Neurite Tracer plugin.

RNA interference

F11 cells were transfected with Dharmacon siGENOME siRNA *SMART*pools (Supplementary Table S6) using the DharmaFECT 3 transfection reagent as previously described (34). Neurite outgrowth was quantified two days later using a Cellomics ArrayScan HCS Reader as decribed above. Experiments were replicated at least three times and representative data were selected for representation. For statistical analysis, neurite lengths were first normalized against untreated control cells (within plate normalization) in order to compare experiments over time (between plates). Next, one-way ANOVA followed by a Dunett's *post hoc* test was used to determine significance against control siRNA-transfected cells.

ChIP and quantitative (RT–)PCR analysis

F11 cells were plated in 15-cm plates, and stimulated with 10 μM forskolin and 10 μM ciglitazone or DMSO as control for 24 h. Chromatin of F11 cells was then cross-linked with 1% formaldehyde for 10 min and subsequently quenched with 125 mM glycine for 5 min. Cells were washed with cold PBS, nuclei were extracted with cell lysis buffer (10 mM EDTA, 10 mM HEPES, 0.25% Triton X-100) and lysed with SDS lysis buffer (1% SDS, 10 mM EDTA in 20 mM Tris–HCl). Cross-linked chromatin was sheared with four pulses of 15 s yielding products of 200–1000 bp in length. Immunoprecipitation was performed with anti-PPARγ (H-100, Santa Cruz Biotechnology) overnight with rotation at 4°C. Immuno-complexes were then captured with protein A/G beads (Santa Cruz Biotechnology) pre-incubated with sonicated salmon sperm DNA. Complexes were washed and eluted with elution buffer (1% SDS, 100 mM $NaHCO_3$). The eluates were proteinase K treated (215 μg/ml) and incubated at 65°C for overnight. DNA was purified by phenol/chloroform isolation and subsequent ethanol precipitation. Quantitative PCR was performed using site-specific primers in duplicate on a Roche LightCycler with 2× SYBR green ready reaction mix (Applied Biosystems). Normalized enrichment values were calculated by subtracting the *Ct* value of the IP sample from the *Ct* value of the mock IP samples, each normalized to the input sample. Promoter regions with >1.5 log-fold enrichment were considered as true targets. For gene expression level measurements, RNA was isolated from F11 cells using Trizol and reverse-transcribed into cDNA as previously described (29). *Ct* values were normalized to *Gapdh* and *Nse* as reference genes. Fold changes were calculated relative to DMSO-treated cells. Specificity of all primers was checked by visual inspection of dissociation curves.

RESULTS

LLM3D: a methodological and statistical improvement in gene set enrichment analysis

Input to the main statistical analysis in LLM3D is a defined set of gene expression clusters, TFBSs and GO terms. For each TFBS–GO pair of interest, LLM3D cross-classifies all genes according to GO annotation, TFBS presence, and gene expression to obtain a 3D contingency table (Figure 1C). The main statistical analysis of LLM3D consists of finding the best model that describes the observed counts in this table. The most basic model, i.e. the model that assumes that gene expression, GO annotation and TFBS presence are mutually independent, is referred to as the null model ($M^{(0)}$). The underlying hypothesis of mutual independence is tested using a likelihood ratio test statistic (20). When this hypothesis is rejected, LLM3D considers eight alternative models that differ in the dependence relationships they imply between gene expression, GO annotation and TFBS presence, and then selects the best model using the Akaike information criterion (AIC) (21). Models $M^{(1)}$, $M^{(2)}$ and $M^{(3)}$ all predict that one variable is independent of the other two. For instance, model $M^{(1)}$ predicts that gene expression and GO variables are dependent, but that TFBS presence is independent of gene expression and GO. Whenever LLM3D selects one of these three models, we conclude that there is no functional interaction between TFBS, gene expression and gene function. Models $M^{(4)}$, $M^{(5)}$, $M^{(6)}$, $M^{(7)}$ and $M^{(s)}$ on the other hand all imply mutual dependencies between gene expression, GO annotation and TFBS presence (see Supplementary Methods for details), and are used by LLM3D to predict functional gene regulatory interactions. LLM3D next calculates for each predicted TFBS the relative enrichment of target genes in the different clusters using a model-based score that is used both to rank and visualize the results of the analysis. Because LLM3D considers all three variables jointly, we expect it to perform better in comparison with existing enrichment-based methods. A detailed description of LLM3D is provided in the Supplementary Methods. LLM3D is available as an R-package.

Identification of functional gene regulatory interactions in yeast

We used both LLM3D and MGSI to predict gene regulatory interactions controlling the yeast metabolic cycle. It is estimated that approximately half of all yeast genes show periodic expression during the metabolic cycle. These genes can be divided into three large expression

clusters of tightly co-regulated genes: oxidative (Ox; 1023 genes), reductive/building (Rb; 977 genes) cluster, and reductive/charging (Rc; 1510 genes) (25) (Figure 2A). The principal difference between LLM3D and MGSI is depicted in Figure 2B. In the given example, ACE2 binding sites are not detected in association with 'mitosis' genes in the three yeast metabolic cycle expression clusters separately by MGSI, whereas LLM3D reveals a significant association of ACE2 binding sites with 'mitosis' genes considering all expression clusters simultaneously. The enrichment of ACE2/'mitosis' genes is highest in the Rb cluster, which is consistent with the fact that cell division is initiated during the Rb phase of the cycle (Figure 2A).

To compare MGSI and LLM3D performance, we only considered the top-20 TFs for which both methods predicted significant TF–target gene associations. We next used *in vivo* yeast TF–target gene interactions reported in the MacIsaac refined regulatory map (MRM) (26) to determine whether these predictions are either true or false. Under the assumption that MRM contains true TF–target gene interactions, predictive performance can be evaluated using the performance quality statistics proposed by Tompa *et al*. (12). We calculated the sensitivity (Sn), the correlation coefficient (CC) and average site performance (ASP), both for MGSI and for LLM3D. The ASP statistic in particular summarizes the overall quality of the predictions, and provides a good measure of predictive performance.

On average, LLM3D achieved a considerable increase in predictive performance compared with MGSI at an equivalent FDR of 10% (Figure 2C–E). LLM3D showed increased Sn values, even when the original and less stringent MGSI *P*-values were used ($P < 0.05$) without correction for multiple testing (Figure 2C). This increase in sensitivity did not simply result from an overall increase in TF–target gene predictions, as the observed average ASP values were also consistently higher [0.34 for LLM3D compared to only 0.24 ($q < 0.1$) or 0.27 ($P < 0.05$) for MGSI; Figure 2E], indicating that LLM3D predictions have higher true/false positive ratios than MGSI predictions. In addition, a modest increase in CC values was observed compared with MGSI predictions, only when the latter were appropriately corrected for multiple testing (Figure 2D). A more detailed evaluation of predictive performance for all 20 TFs tested is provided in Supplementary Table S1. Results reported in Figure 2 and in Supplementary Table S1 were obtained using MRM as the repository of true interactions. Similar results were obtained using the YEASTRACT bibliographic repository of true TF–target gene interactions (27) (Supplementary Table S2). Thus, using two different sources of true TF–target gene interactions, LLM3D provides a significant improvement in predictive performance compared with MGSI.

To evaluate LLM3D performance from a functional perspective, we next selected for each expression cluster the four TFs that have the largest contribution to the increase in true positive TF–target gene predictions compared with MGSI. In addition to well-known key regulators of the metabolic cycle (e.g. SKN7 and FKH1) (35), this list contains several potential novel cluster-specific regulators, such as RAP1, FHL1, HAP1, REB1 and UME6 (Figure 2F). Importantly, the GO terms that are associated with these TFs all correspond with the biological processes that characterize each cluster (Figure 2A). Interestingly, both RAP1 and FHL1 were recently implicated in growth-rate dependent changes in ribosome synthesis (36), which confirms their predicted relative importance in the Ox cluster. In conclusion, LLM3D achieves a significant gain in statistical power and improved predictive performance compared with existing methods with respect to the detection of functional TF–target gene interactions in yeast.

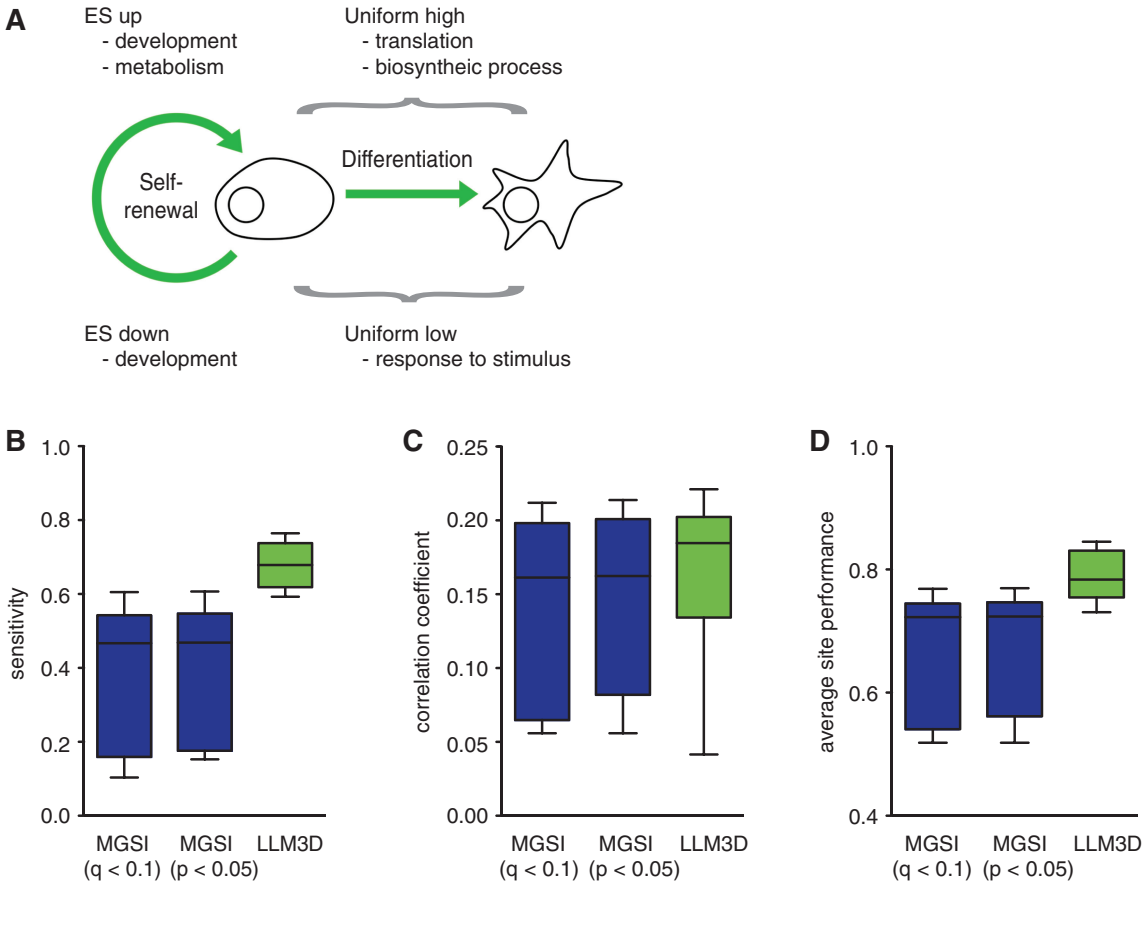
Identification of functional gene regulatory interactions in mouse ES cells

We next wanted to test LLM3D performance on mammalian gene expression data. We used both LLM3D and MGSI to predict the transcriptional regulation of genes that are involved in controlling self-renewal and differentiation of mouse ES cells. Ouyang and colleagues (28) defined four clusters of genes with characteristic expression patterns in ES cells. These clusters are based on combined RNA-Seq data (37) and gene expression microarray data (38). The first two clusters include genes that are either induced (uniform high; 668 genes) or repressed (uniform low; 838 genes) in both ES cells and differentiated cells. The other two clusters include genes that are either upregulated (Es up; 782 genes) or downregulated (Es down; 831 genes) in ES cells only (Figure 3A). We restricted our analysis to 12 TFs that are known to play a role in ES cell biology and for which genome-wide ChIP-Seq binding profiles are available (23). This allowed us to define true targets and to benchmark LLM3D predictive performance as we did for the yeast metabolic cycle data.

Again, LLM3D showed a significant increase in true positive predictions and in Sn, CC and ASP values (Figure 3B–D; Supplementary Table S3). Importantly, LLM3D predicted a significant role for two key regulators of ES cell self-renewal, Oct4 and Nanog, whereas MGSI did not. In addition to Oct4 and Nanog, the highest increase in true positive predictions was observed for Esrrb. For these three TFs, LLM3D predicted the same cluster-specific GO associations (Figure 3E). These GO terms indeed reflect the cluster-specific biological processes identified by Ouyang *et al*. (28), i.e. 'development' and 'morphogenesis' in the ES up and ES down clusters, 'translation' and 'metabolism' in the uniform high cluster, and 'response to stimulus' in the uniform low cluster (Figure 3A). Our findings corroborate earlier reports that Oct4, Nanog and Esrrb regulate ES cell gene expression in a combinatorial manner, and that they can either activate or repress genes depending on the context (28). Thus, LLM3D provides improved detection of functional gene regulatory interactions in mammalian gene expression data.

Identification of transcriptional regulators of neuronal regeneration

We next used LLM3D to predict transcriptional regulatory interactions underlying neuronal regeneration.



E

TF	Cluster	Associated GO-terms
Oct4 Nanog Esrrb	ES up / ES down	ribosome biogenesis, rRNA metabolic process, RNA processing, regulation of translation
	Uniform high	mitochondrion organization and biogenesis, energy derivation by oxidation of organic compounds, oxidation reduction, phosphate metabolic process, post-translational protein modification
	Uniform low	response to chemical stimulus, cellular protein catabolic process, proteolysis, post-translational protein modification, protein ubiquitination

Figure 3. LLM3D correctly infers gene regulatory interactions from mouse ES cell self-renewal gene expression data. (**A**) Schematic representation of ES cell self-renewal and differentiation. Indicated are the four gene expression clusters (ES up, ES down, uniform high and uniform low) identified by Ouyang *et al.* (28) together with the biological processes that dominate each cluster. (**B**) Sensitivity (Sn) of LLM3D and MGSI with respect to predicting transcriptional regulators in mouse ES cells. LLM3D shows higher Sn values compared with MGSI, even when the stringency of the latter is reduced to a *P*-value cut-off of 0.05 without correction for multiple testing. (**C**) Correlation coefficient (CC) for LLM3D and MGSI predictive performances. (**D**) Average site performance (ASP) of LLM3D and MGSI. (**E**) In contrast to MGSI, LLM3D correctly inferred a role for Nanog and Oct4 and provided a significant improvement of Essrb target gene predictions. The associated GO terms overlap for the three TFs but differ per expression cluster.

We first generated genome-wide expression profiles of dorsal root ganglion (DRG) neurons following nerve damage (Figure 4A). DRG neurons extend one peripheral axon into the spinal nerve and one central axon into the DR. The peripheral axon regenerates vigorously; while in contrast, the central axon has little regenerative capacity. For this study, two groups of animals were subjected either to SN or DR crush, and at 12, 24, 72 h and 7 days after the crush, lumbar DRGs L4, L5 and L6 were dissected and total RNA was extracted. Samples were then processed and hybridized to Agilent 44K rat whole-genome arrays. Bayesian analysis of time-series

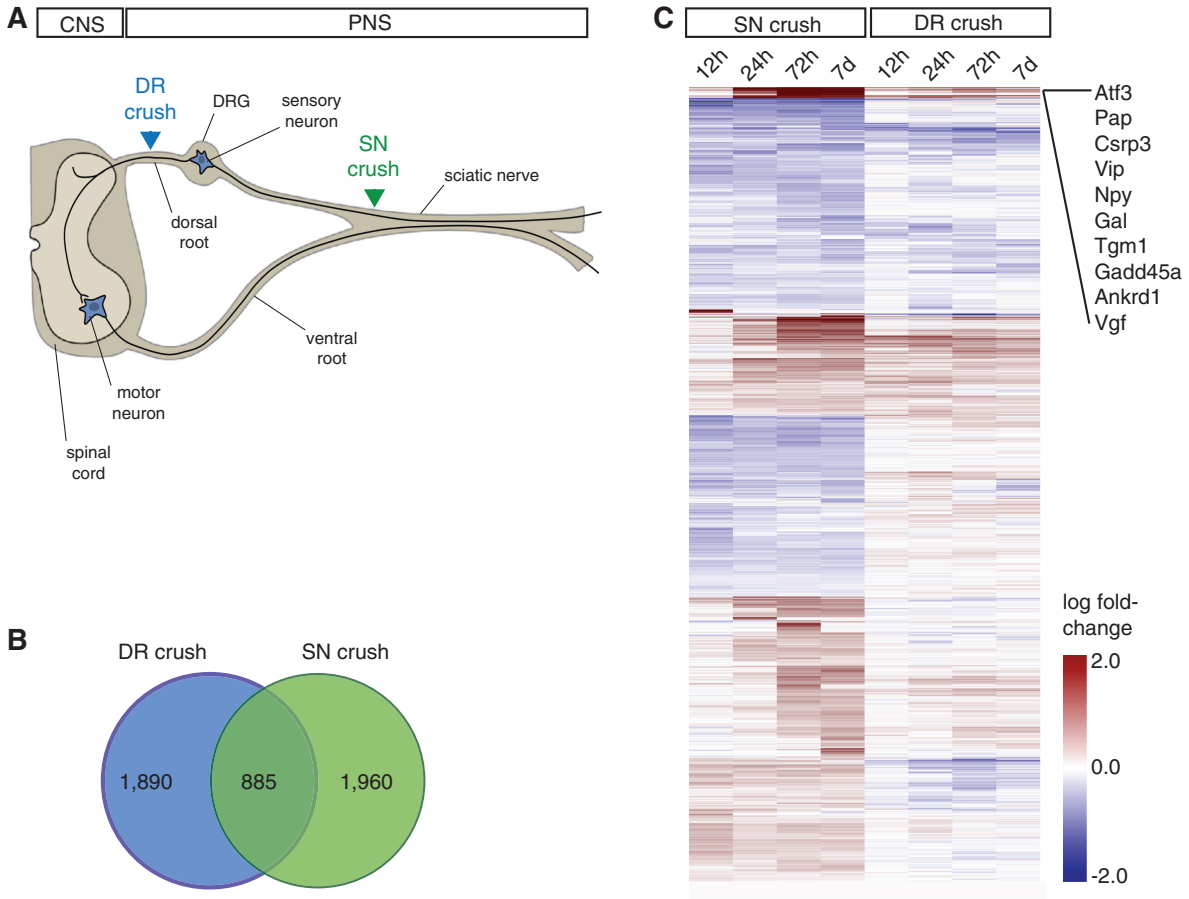


Figure 4. Genome-wide identification of regeneration-associated genes. (**A**) Schematic representation of the sensory-motor neuron circuitry and the location of the DRG. A dorsal root crush injures the central projections of DRG sensory neurons, whereas a peripheral nerve crush injures the peripheral projections of the same neurons. (**B**) Venn diagram showing the number of significantly regulated genes identified in each crush paradigm. The relatively small overlap indicates that SN and DR crush elicit distinct gene responses in DRG neurons. (**C**) Heat map showing expression profiles of all significantly regulated genes after SN or DR crush.

(BATS) (32,33) was used to identify 2845 genes that are significantly regulated after SN crush and 2775 genes that are significantly regulated after DR crush relative to control (Supplementary Table S4; GEO accession number GSE21007). Out of the 4735 regulated genes in total, only 885 genes were regulated in both crush paradigms and 3850 were regulated in a paradigm-specific manner (Figure 4B), which confirms the notion that SN and DR crush elicit very distinct gene expression responses in DRG neurons (29). In line with previous gene expression studies (29,39–41), we find a strong and SN crush-specific up-regulation of regeneration-associated genes, including *Atf3*, *Pap*, *Vip*, *Npy*, *Gal*, *Tgm1*, *Csrp3*, *Ankrd1*, *Gadd45a* and *Vgf* (Figure 4C; Supplementary Table S4).

We separated all 4735 regulated genes into two distinct expression clusters: one cluster of genes that are higher expressed after SN crush than after DR crush (named SN > DR), and one cluster of genes that are higher expressed after DR crush than after SN crush (named DR > SN) (Figure 5A). We next used LLM3D to predict transcription regulatory interactions underlying each gene expression cluster. After correction for multiple testing, predicted TFBSs were ranked based on cluster-specific enrichment, and the 50 TFBSs with the highest gene cluster-specific regulatory potential were selected. These 50 TFBSs include 27 that were only identified by LLM3D, and not by MGSI (Figure 5B).

In the absence of a repository for true positive TF–target gene interactions, we decided to test whether any of the TFBSs that were identified exclusively by LLM3D could reflect functional gene regulatory interactions in the context of regenerative axon growth. We used F11 cells as an *in vitro* model. F11 cells are neuroblastoma cells derived from rat embryonic DRG neurons (42). They express many DRG neuron markers (43,44) and display cAMP-induced neurite outgrowth (45). We previously showed that DRG regeneration-associated TFs also are important for F11 neurite outgrowth (29,34). For the 27 TFBSs that were exclusively identified by LLM3D, we could unequivocally identify 18 corresponding rat TFs. RNAi-mediated knockdown of eight of these TFs in F11 cells resulted in a significant increase or decrease in neurite outgrowth (Figure 6). These findings demonstrate that LLM3D identified functional gene regulatory interactions that remained undetected by MGSI.

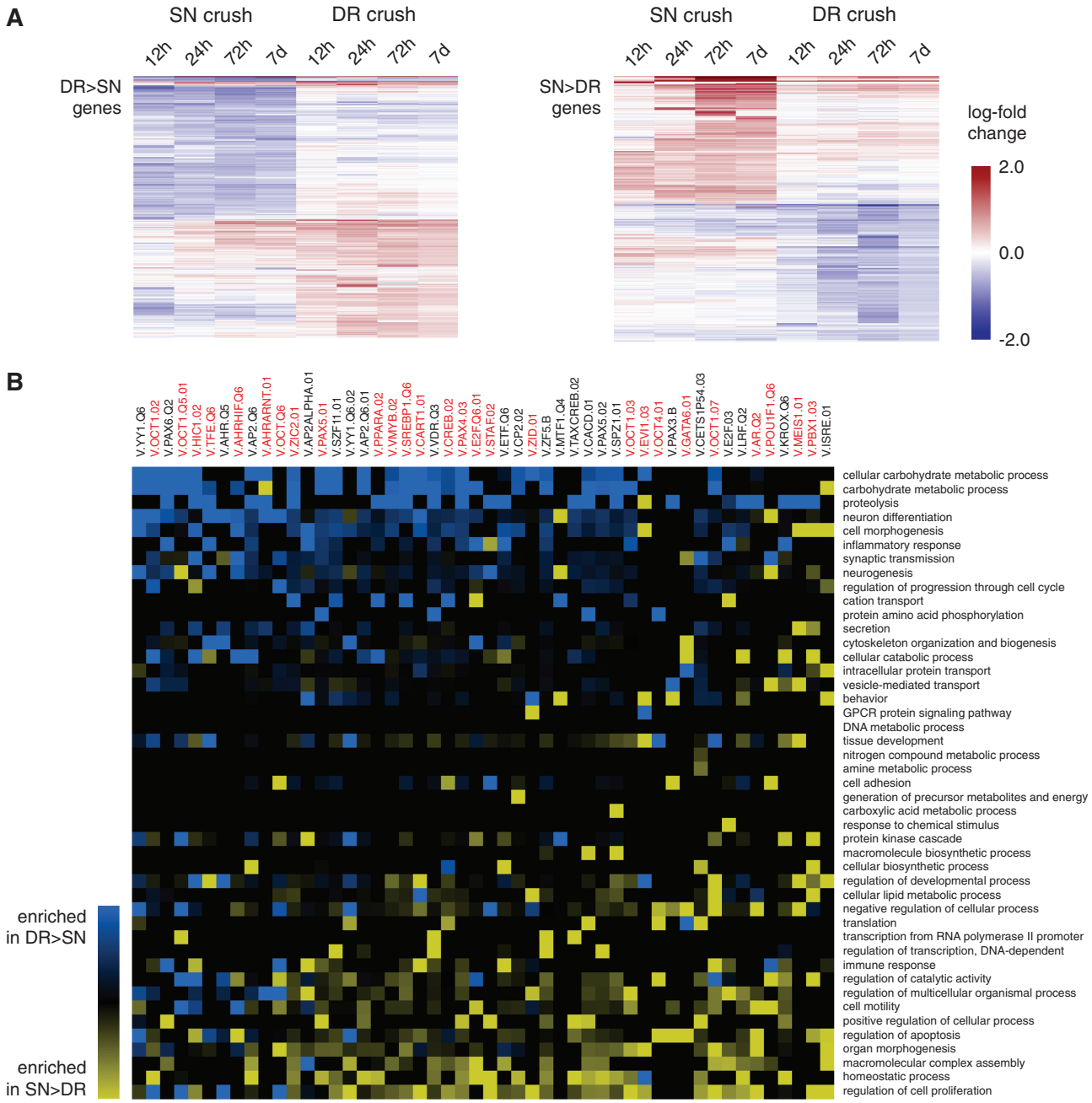


Figure 5. LLM3D identifies more gene regulatory interactions in regeneration-associated genes than MGSI. (**A**) All genes that are significantly regulated after SN or DR injury were subdivided into two clusters: DR > SN and SN > DR (see text for details). (**B**) Heat map showing the top-50 TFBS–GO associations detected in DR > SN genes (blue) and in SN > DR genes (yellow). TRANSFAC binding sites are on the horizontal axis (binding sites indicated in red were only detected with LLM3D, binding sites indicated in black were also identified with MGSI); GO terms are on the vertical axis.

PPARγ is a repressor of neuronal regeneration-associated genes

Because PPAR binding sites were detected in 'neuron differentiation' GO genes (Figure 5B) and knockdown of PPARγ strongly and significantly reduced neuronal differentiation of F11 cells (Figure 6), we decided to study the role of PPAR TFs in neuronal regeneration in more detail. Most predicted PPAR target genes are expressed in neurons (Supplementary Table S5), which suggests that PPAR TFs may enhance regeneration by regulating the expression of neuron-intrinsic regeneration-associated genes. Therefore, we tested the effects of different PPAR agonists and antagonists on neurite outgrowth from F11 cells and from primary adult DRG neurons. Stimulation of PPARγ, but not PPARα, stimulated neurite outgrowth from primary DRG neurons and from F11 cells, whereas blocking PPARγ, but not PPARα, inhibited neurite outgrowth in both cell types (Figure 7A–D). These findings

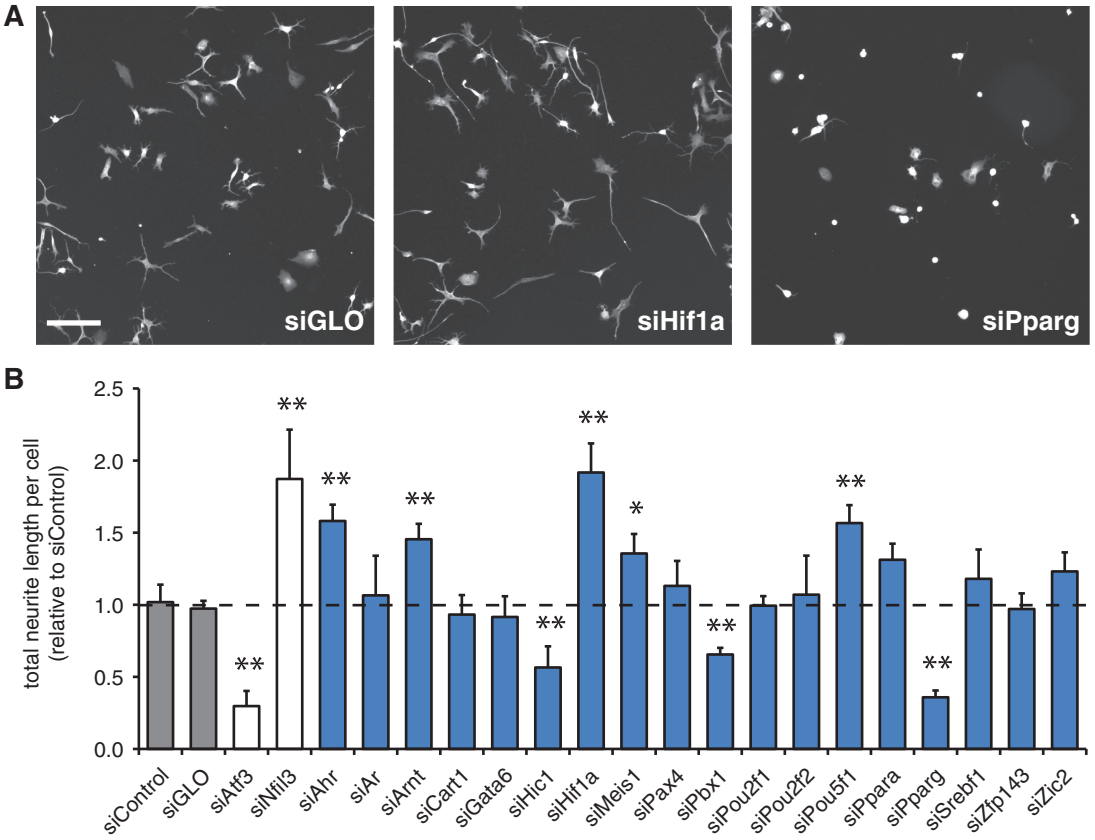


Figure 6. LLM3D-predicted transcriptional regulators of regeneration-associated genes are functional in neurite outgrowth assays *in vitro*. (**A**) Example images of F11 cells transfected with control siRNAs (left), siRNAs against *Hif1a* (middle), or siRNAs against *Pparg* (right). Scale bar: 200 μm. (**B**) siRNA-mediated knockdown of TF expression in F11 cells shows that eight out of 18 TFs predicted by LLM3D but not by MGSI significantly increase or decrease forskolin-induced neurite outgrowth (one-way ANOVA: $F_{21,109} = 34.3$, $P < 0.001$). Representative data from three independent experiments are shown. Grey and white bars represent negative and positive controls respectively. Bars represent mean neurite total length per cell ± SD; $^{**}P < 0.01$; $^{*}P < 0.05$ (Dunett's *post hoc* test).

show that activation of PPARγ in primary adult DRG neurons, which closely resemble the *in vivo* DRG regeneration paradigm (29,34), stimulates neurite outgrowth. Primary DRG cultures are however mixed neuron/glia cultures, and the effects of PPARγ activation or inhibition on DRG neuron outgrowth might be indirectly mediated by glial cells. F11 cell cultures on the other hand contain no glia, and the fact that we could replicate our results in F11 cells indicates that PPARγ is a neuron-intrinsic stimulator of neurite outgrowth.

To test whether PPARγ binds directly to the promoters of predicted target genes, we next performed quantitative ChIP. F11 cells were stimulated with the PPARγ agonist ciglitazone or with DMSO (control) and chromatin complexes were cross-linked after 24 h and subjected to ChIP using an antibody specific for PPARγ. Immunoprecipitated DNA was then analyzed using quantitative PCR. PCR primers were designed to recognize ~100 bp promoter regions containing the predicted PPAR binding sites for nine randomly chosen predicted target genes. As negative controls we used primers recognizing promoter regions of *Icer* and *JunD* that lack PPAR binding sites. For seven promoter regions tested, we found a specific interaction with PPARγ, which in most cases was further induced by ciglitazone (Figure 7E). These findings indicate that LLM3D predicts within a given functional context (i.e. neuron differentiation) PPARγ target gene interactions with an accuracy of >75%.

We finally measured the effect of ciglitazone on the expression of the five predicted target genes that show the highest PPARγ binding. Quantitative PCR measurements indicate that activation of PPARγ with ciglitazone significantly reduces the expression of three of these genes (Figure 7F), which demonstrates that PPARγ acts as a ligand-dependent repressor of gene expression. Importantly, PPARα agonist Wy-14643 did not affect gene expression levels, nor did any of the pharmacons affect the expression levels of PPARγ (Figure 7F).

DISCUSSION

Reverse engineering transcriptional regulatory networks from experimental data presents great challenges, particularly in higher organisms. As more genome-wide gene expression and functional data sets become available, there is a growing need for computational methods to analyze these data and accurately infer regulatory relationships from them. Of particular interest are those methods that

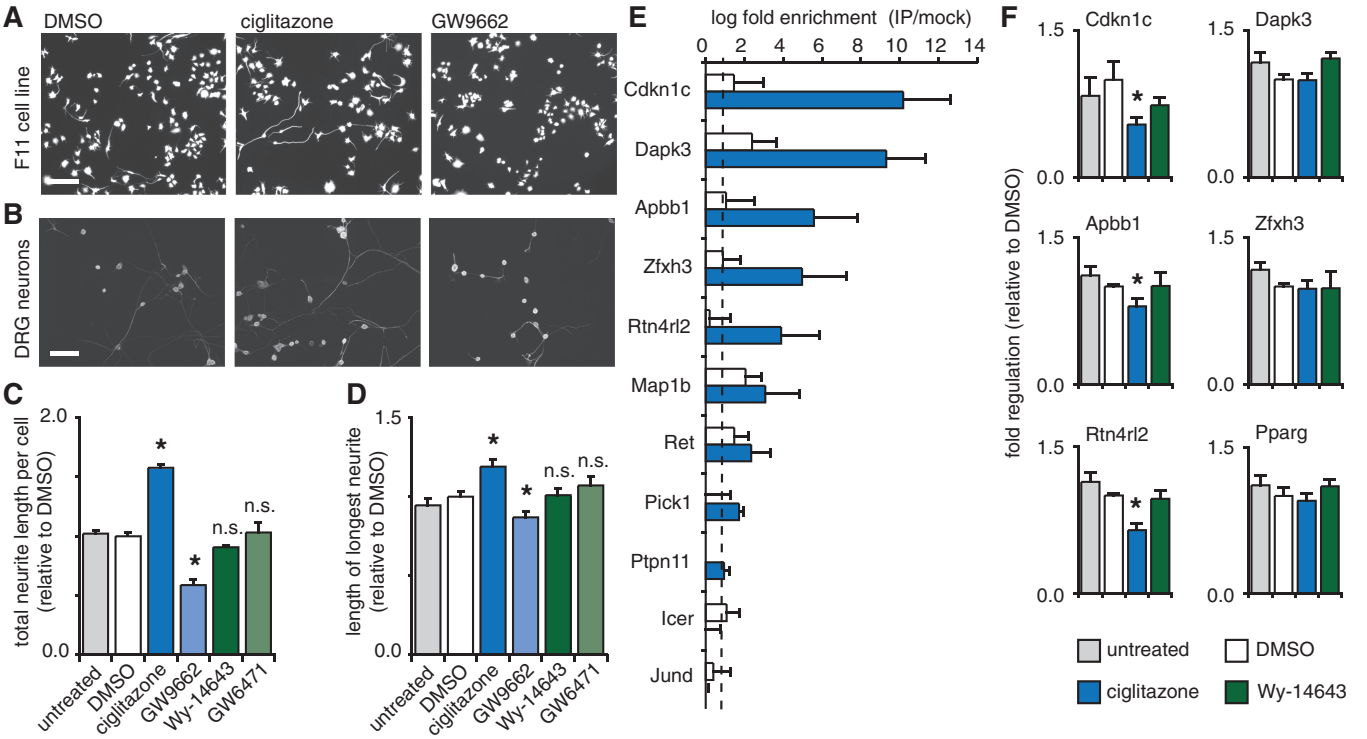


Figure 7. Experimental validation of PPARγ binding sites in regeneration-associated genes. (**A**) F11 cells treated with PPARγ agonist ciglitazone show increased neurite outgrowth, whereas cells treated with PPARγ antagonists GW9662 show decreased neurite outgrowth. Scale bar: 200 μm. (**B**) Similar results were obtained for cultured primary adult DRG neurons. Scale bar: 200 μm. (**C**) Quantification of the effects of ciglitazone and GW9662 on F11 cell neurite outgrowth. (**D**) Quantification of the effects of ciglitazone and GW9662 on primary DRG neurite outgrowth. Note that PPARα agonist Wy-14643 and antagonist GW6471 do not affect neurite outgrowth. (**E**) PPARγ binds to the promoters of predicted 'neuron differentiation' target genes. Anti-PPARγ immunoprecipitated chromatin from F11 cells treated with DMSO (negative control; white bars) or PPARγ agonist ciglitazone (blue bars) was quantified by PCR using site-specific primers. *Icer* and *JunD* were included as negative control genes. All predicted target genes tested, except *Pick1* and *Ptpn11*, show PPARγ binding above background (dashed line), and for most genes this binding was strongly enhanced by ciglitazone. (**F**) Three out of the top-5 PPARγ-binding genes show a significant reduction in expression after ciglitazone treatment (blue bars) compared with DMSO treatment (white bars). PPARα agonist Wy-14643 did not affect gene expression levels (green bars), nor did ciglitazone or Wy-14643 affect the expression levels of PPARγ. Bars represent means ± SD; $*P < 0.01$; n.s., not siginificant.

automatically generate experimentally testable hypotheses regarding the direct regulation of genes by DNA binding TFs. Combining heterogeneous sources of information, including genome-wide gene expression data, DNA sequence information and functional annotation, may prove to be essential to accurately predict true regulatory relationships. Indeed, current methods that allow the integration of gene expression data, TFBS motifs and ChIP data predict TF–target gene interactions with high accuracy (46-48). For mammalian systems, however, high quality experimental TF binding data is often lacking or available for a few TFs only. Here, we present a new method that can be used when no experimental TF binding data is available, and that offers a significant improvement over currently used enrichment-based methods. We show that our method can be applied to predict novel, condition-specific sets of transcriptional targets in the context of the complexity of the mammalian genome.

The main limitation of existing methods is that they do not model the joint dependence between gene expression, TFBS presence and gene function. SEA-based methods for instance produce lists in which enriched TFBS and GO terms occur separately. From such lists it is unclear how GO terms and TFBS are jointly related to the gene sets of interest, and thus it is not possible to directly use SEA results to predict functionally homogenous sets of TF target genes. MGSI-based methods on the other hand try to circumvent this problem by using pre-defined GO-expression gene sets, and subsequently test these sets for enrichment of TFBSs. Although it makes sense to search for TFBS enrichment in functionally homogeneous sets of co-expressed genes, there are important conceptual problems with this approach that compromise the analysis and adversely affect the power to detect biologically meaningful associations. For instance, MGSI does not really consider gene expression, TFBS presence and GO annotation jointly, but rather collapses gene expression and GO annotation into a single combined variable before computational analysis. Thus, important information about the joint dependence of all three variables is lost. Moreover, by analyzing multiple disjoint gene expression clusters, MGSI aggravates the multiple-testing problem because separate tests are performed for each cluster. LLM3D efficiently deals with both problems; it allows modeling of the joint distribution between all variables and reduces the number of tests to be performed.

We validated LLM3D performance using published yeast and mammalian gene expression and TF binding data sets. In yeast metabolic cycle gene expression clusters, LLM3D detects experimentally validated TFBSs that remain undetected using MGSI. Moreover, for most of these TFBSs, the true positive target gene prediction rates are significantly higher than found with MGSI. A similar increase in performance of LLM3D compared with MGSI was observed for mouse ES cell gene expression data. Although true positive rates did not increase as much as in the yeast example, LLM3D was uniquely able to identify target gene interactions for classical key regulators of the cell cycle (i.e. Nanog and Oct4) and showed significantly improved target gene detection for several other TFs (e.g. Esrrb, E2f1 and Stat3). Importantly, in both the yeast metabolic cycle data and the mouse ES cell data, LLM3D identified known and novel TFs in association with GO terms that reflect the biological processes underlying each expression cluster. We conclude that LLM3D not only provides a significant computational improvement over MGSI, but it also detects biologically relevant TF–target gene interactions, both in yeast and in mammals.

Although, we demonstrate improved predictive performance for LLM3D compared with MGSI, we also noticed a significant increase in false-positive predictions. This is most likely due to the fact that LLM3D is able to detect TFBS enrichments originating from weak TF–target gene interactions that are surrounded by noise. To be able to increase specificity, we have implemented the possibility to restrict LLM3D analysis to human/mouse/rat (HMR) conserved binding sites. Also, a *post hoc* procedure has been included to select for TF–target gene interactions that show the highest gene cluster-specific enrichment. Both options may be used to decrease false positive levels and consequently improve the specificity of LLM3D.

We next used LLM3D to identify gene regulatory interactions underlying neuronal regeneration. We first used microarray analysis to define two clusters of genes with differential expression in DRGs during either successful or unsuccessful axonal regeneration. Out of 50 LLM3D-predicted TFBSs that showed the highest gene cluster-specific enrichment, 27 were identified exclusively by LLM3D. For these 27 TFBSs, we could identify 18 corresponding rat TFs, 8 of which significantly increased or decreased neurite outgrowth after siRNA-mediated knockdown in F11 cells. Most notably, knockdown of AHR, ARNT and HIF1α, which are structurally related bHLH TFs involved in the cellular response to hypoxia (49), all strongly enhanced neurite outgrowth, whereas knockdown of PBX1, HIC1 and PPARγ strongly reduced neurite outgrowth. Thus, LLM3D identified potential novel TFs and transcriptional regulatory pathways involved in neurite outgrowth.

We decided to focus on one of these newly identified TFs, i.e. PPARγ. Recent work showed that PPARγ is expressed in several neuronal cell lines and may promote differentiation and neurite outgrowth (50,51). Moreover, activation of PPARγ in spinal cord injury models has beneficial effects on the functional outcome (52,53), but it is not clear whether these effects are directly on the damaged neurons, or whether PPARγ reduces the secondary inflammatory response (54). Our results add to these findings, and show that PPARγ, but not PPARα, stimulates neurite outgrowth of DRG neurons. Moreover, this effect of PPARγ is neuron-intrinsic since we also observe it in DRG-like F11 cells, which in the presence of forskolin acquire a neuronal phenotype. Activated PPARγ binds to promoters of predicted target genes and reduces their expression. Importantly, several predicted PPARγ target genes are known inhibitors of neurite outgrowth (e.g. *Rtn4rl2*, *Slit1*, *Hes5*; see Supplementary Table S5), which suggests that PPARγ promotes neurite outgrowth by repressing growth-inhibitory genes. At this moment we can only speculate about the relevance of these findings for neuronal regeneration *in vivo*. The primary ligands of PPARγ are polyunsaturated fatty acids (55). Following nerve crush and degeneration of the myelin sheath, free myelin lipids are taken up by macrophages and released again as fatty acids to be incorporated into the newly forming myelin sheath (56). Injured axons might benefit from fatty acid production in the damaged nerve, and the neuron-intrinsic lipid sensing properties of PPARγ may play an important role in conveying injury signals from the crush site to the nucleus. This hypothesis is supported by several reports showing beneficial effects of fatty acids on neurite outgrowth *in vitro* (57,58) and on neuronal regeneration *in vivo* (52,53), and the induction of fatty acid binding proteins in regenerating axons (59).

One of the challenges left unaddressed in the current implementation of our method is that transcriptional regulation in higher organisms is believed to be highly combinatorial, and that the spatiotemporal expression of genes is influenced by multiple regulatory TFs that form complexes at multiple TFBSs. Although some basic models for the cooperative effect of multiple TFs on the expression of target genes have been suggested (10,35,60,61), in general the *cis*-regulatory grammar underlying gene regulation is still poorly understood. Moreover, combinatorial models of gene regulation are difficult to validate and the effect of different TFs on target genes is therefore most often studied independently. As soon as reliable and genome-wide descriptions of *cis*-regulatory modules become available, this information can easily be incorporated into LLM3D to allow modeling of *cis*-regulatory modules in addition to individual TFBSs. For instance, Robertson *et al*. (62) described a database system (cisRED) that currently allows genome-scale mapping of small regulatory modules. Binary predictors for the presence or absence of such regulatory modules can be used by LLM3D in the same way as we used TRANSFAC motifs in the present study.

In conclusion, LLM3D provides an important improvement over existing computational methods in identifying functional TFBSs from gene expression data. Its unique property of testing the joint association between multiple features (e.g. gene expression, gene function and TFBS occurrence) based on one table allows further generalization to tables with more dimensions including additional relevant gene attributes. The implementation of such

multidimensional computational methods will be of critical importance in order to extract biologically meaningful information from the increasing number, size and diversity of data sets generated by biologists.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the microarray facility team at the VU University Medical Center, in particular Paul P. Eijk, Danielle Israeli and Bauke Ylstra, for technical support.

FUNDING

The Netherlands Organization for Scientific Research (CLS grant 635.100.008 to M.C.M.D.G.); the Dutch Ministry of Economic Affairs (SenterNovem grant ISO52022 to J.V.); and the Center for Medical Systems Biology (CMSB) in the framework of the Netherlands Genomics Initiative (NGI).

Conflict of interest statement. None declared.

REFERENCES

1. Aerts,S., Thijs,G., Coessens,B., Staes,M., Moreau,Y. and De Moor,B. (2003) Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res.*, **31**, 1753–1764.
2. Cui,X., Wang,T., Chen,H.S., Busov,V. and Wei,H. (2010) TF-finder: a software package for identifying transcription factors involved in biological processes using microarray data and existing knowledge base. *BMC Bioinformatics*, **11**, 425.
3. Ernst,J., Plasterer,H.L., Simon,I. and Bar-Joseph,Z. (2010) Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res.*, **20**, 526–536.
4. Hannenhalli,S. (2008) Eukaryotic transcription factor binding sites–modeling and integrative search methods. *Bioinformatics*, **24**, 1325–1331.
5. Kel,A.E., Gossling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.
6. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.
7. Won,K.J., Agarwal,S., Shen,L., Shoemaker,R., Ren,B. and Wang,W. (2009) An integrated approach to identifying cis-regulatory modules in the human genome. *PLoS ONE*, **4**, e5501.
8. Roider,H.G., Kanhere,A., Manke,T. and Vingron,M. (2007) Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, **23**, 134–141.
9. Ward,L.D. and Bussemaker,H.J. (2008) Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. *Bioinformatics*, **24**, i165–171.
10. Warner,J.B., Philippakis,A.A., Jaeger,S.A., He,F.S., Lin,J. and Bulyk,M.L. (2008) Systematic identification of mammalian regulatory motifs' target genes and functions. *Nat. Methods*, **5**, 347–353.
11. Zinzen,R.P., Girardot,C., Gagneur,J., Braun,M. and Furlong,E.E. (2009) Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, **462**, 65–70.
12. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y., Kent,W.J. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.
13. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
14. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
15. Nam,D. and Kim,S.Y. (2008) Gene-set approach for expression pattern analysis. *Brief Bioinform.*, **9**, 189–197.
16. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
17. Reimand,J., Kull,M., Peterson,H., Hansen,J. and Vilo,J. (2007) g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.*, **35**, W193–W200.
18. Carmona-Saez,P., Chagoyen,M., Tirado,F., Carazo,J.M. and Pascual-Montano,A. (2007) GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol.*, **8**, R3.
19. Nogales-Cadenas,R., Carmona-Saez,P., Vazquez,M., Vicente,C., Yang,X., Tirado,F., Carazo,J.M. and Pascual-Montano,A. (2009) GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nucleic Acids Res.*, **37**, W317–W322.
20. Christensen,R. (1997) *Log-Linear Models and Logistic Regression*, 2nd edn. Springer, Berlin.
21. Akaike,H. (1974) A new look at the statistical model identification. *IEEE Trans. Autom. Control*, **19**, 716–723.
22. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
23. Chen,X., Xu,H., Yuan,P., Fang,F., Huss,M., Vega,V.B., Wong,E., Orlov,Y.L., Zhang,W., Jiang,J. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.
24. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
25. Tu,B.P., Kudlicki,A., Rowicka,M. and McKnight,S.L. (2005) Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science*, **310**, 1152–1158.
26. MacIsaac,K.D., Wang,T., Gordon,D.B., Gifford,D.K., Stormo,G.D. and Fraenkel,E. (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, **7**, 113.
27. Teixeira,M.C., Monteiro,P., Jain,P., Tenreiro,S., Fernandes,A.R., Mira,N.P., Alenquer,M., Freitas,A.T., Oliveira,A.L. and Sa-Correia,I. (2006) The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Res.*, **34**, D446–D451.
28. Ouyang,Z., Zhou,Q. and Wong,W.H. (2009) ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc. Natl Acad. Sci. USA*, **106**, 21521–21526.
29. Stam,F.J., MacGillavry,H.D., Armstrong,N.J., de Gunst,M.C., Zhang,Y., van Kesteren,R.E., Smit,A.B. and Verhaagen,J. (2007) Identification of candidate transcriptional modulators involved in successful regeneration after nerve injury. *Eur. J. Neurosci.*, **25**, 3629–3637.
30. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
31. Ritchie,M.E., Silver,J., Oshlack,A., Holmes,M., Diyagama,D., Holloway,A. and Smyth,G.K. (2007) A comparison of

background correction methods for two-colour microarrays. *Bioinformatics*, **23**, 2700–2707.
32. Angelini,C., Cutillo,L., De Canditiis,D., Mutarelli,M. and Pensky,M. (2007) BATS: A Bayesian user-friendly software for analyzing time series microarray data. *Technical Report CNR-IAC 331/07*.
33. Angelini,C., De Canditiis,D., Mutarelli,M. and Pensky,M. (2007) A Bayesian approach to estimation and testing in time-course microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **6**, Article 24.
34. MacGillavry,H.D., Stam,F.J., Sassen,M.M., Kegel,L., Hendriks,W.T., Verhaagen,J., Smit,A.B. and van Kesteren,R.E. (2009) NFIL3 and cAMP response element-binding protein form a transcriptional feedforward loop that controls neuronal regeneration-associated gene expression. *J. Neurosci.*, **29**, 15542–15550.
35. Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I. *et al.* (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 799–804.
36. Fazio,A., Jewett,M.C., Daran-Lapujade,P., Mustacchi,R., Usaite,R., Pronk,J.T., Workman,C.T. and Nielsen,J. (2008) Transcription factor control of growth rate dependent genes in Saccharomyces cerevisiae: a three factor design. *BMC Genomics*, **9**, 341.
37. Cloonan,N., Forrest,A.R., Kolle,G., Gardiner,B.B., Faulkner,G.J., Brown,M.K., Taylor,D.F., Steptoe,A.L., Wani,S., Bethel,G. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods*, **5**, 613–619.
38. Zhou,Q., Chipperfield,H., Melton,D.A. and Wong,W.H. (2007) A gene regulatory network in mouse embryonic stem cells. *Proc. Natl Acad. Sci. USA*, **104**, 16438–16443.
39. Costigan,M., Befort,K., Karchewski,L., Griffin,R.S., D'Urso,D., Allchorne,A., Sitarski,J., Mannion,J.W., Pratt,R.E. and Woolf,C.J. (2002) Replicate high-density rat genome oligonucleotide microarrays reveal hundreds of regulated genes in the dorsal root ganglion after peripheral nerve injury. *BMC Neurosci.*, **3**, 16.
40. Schmitt,A.B., Breuer,S., Liman,J., Buss,A., Schlangen,C., Pech,K., Hol,E.M., Brook,G.A., Noth,J. and Schwaiger,F.W. (2003) Identification of regeneration-associated genes after central and peripheral nerve injury in the adult rat. *BMC Neurosci.*, **4**, 8.
41. Szpara,M.L., Vranizan,K., Tai,Y.C., Goodman,C.S., Speed,T.P. and Ngai,J. (2007) Analysis of gene expression during neurite outgrowth and regeneration. *BMC Neurosci.*, **8**, 100.
42. Platika,D., Boulos,M.H., Baizer,L. and Fishman,M.C. (1985) Neuronal traits of clonal cell lines derived by fusion of dorsal root ganglia neurons with neuroblastoma cells. *Proc. Natl Acad. Sci. USA*, **82**, 3499–3503.
43. Boland,L.M. and Dingledine,R. (1990) Expression of sensory neuron antigens by a dorsal root ganglion cell line, F-11. *Brain Res. Dev. Brain Res.*, **51**, 259–266.
44. Francel,P.C., Harris,K., Smith,M., Fishman,M.C., Dawson,G. and Miller,R.J. (1987) Neurochemical characteristics of a novel dorsal root ganglion X neuroblastoma hybrid cell line, F-11. *J. Neurochem.*, **48**, 1624–1631.
45. Ghil,S.H., Kim,B.J., Lee,Y.D. and Suh-Kim,H. (2000) Neurite outgrowth induced by cyclic AMP can be modulated by the alpha subunit of Go. *J. Neurochem.*, **74**, 151–158.
46. Beyer,A., Workman,C., Hollunder,J., Radke,D., Moller,U., Wilhelm,T. and Ideker,T. (2006) Integrated assessment and prediction of transcription factor binding. *PLoS Comput. Biol.*, **2**, e70.
47. Chen,G., Jensen,S.T. and Stoeckert,C.J. Jr (2007) Clustering of genes into regulons using integrated modeling-COGRIM. *Genome Biol.*, **8**, R4.
48. Youn,A., Reiss,D.J. and Stuetzle,W. (2010) Learning transcriptional networks from the integration of ChIP-chip and expression data in a non-parametric model. *Bioinformatics*, **26**, 1879–1886.
49. Bracken,C.P., Whitelaw,M.L. and Peet,D.J. (2003) The hypoxia-inducible factors: key transcriptional regulators of hypoxic responses. *Cell Mol. Life Sci.*, **60**, 1376–1393.
50. Dill,J., Patel,A.R., Yang,X.L., Bachoo,R., Powell,C.M. and Li,S. (2010) A molecular mechanism for ibuprofen-mediated RhoA inhibition in neurons. *J. Neurosci.*, **30**, 963–972.
51. Miglio,G., Rattazzi,L., Rosa,A.C. and Fantozzi,R. (2009) PPARgamma stimulation promotes neurite outgrowth in SH-SY5Y human neuroblastoma cells. *Neurosci. Lett.*, **454**, 134–138.
52. McTigue,D.M., Tripathi,R., Wei,P. and Lash,A.T. (2007) The PPAR gamma agonist Pioglitazone improves anatomical and locomotor recovery after rodent spinal cord injury. *Exp. Neurol.*, **205**, 396–406.
53. Park,S.W., Yi,J.H., Miranpuri,G., Satriotomo,I., Bowen,K., Resnick,D.K. and Vemuganti,R. (2007) Thiazolidinedione class of peroxisome proliferator-activated receptor gamma agonists prevents neuronal damage, motor dysfunction, myelin loss, neuropathic pain, and inflammation after spinal cord injury in adult rats. *J. Pharmacol. Exp. Ther.*, **320**, 1002–1012.
54. McTigue,D.M. (2008) Potential Therapeutic Targets for PPARgamma after Spinal Cord Injury. *PPAR Res.*, **2008**, 517162.
55. Hihi,A.K., Michalik,L. and Wahli,W. (2002) PPARs: transcriptional effectors of fatty acids and their derivatives. *Cell Mol. Life Sci.*, **59**, 790–798.
56. Goodrum,J.F., Weaver,J.E., Goines,N.D. and Bouldin,T.W. (1995) Fatty acids from degenerating myelin lipids are conserved and reutilized for myelin synthesis during regeneration in peripheral nerve. *J. Neurochem.*, **65**, 1752–1759.
57. Liu,J.W., Almaguel,F.G., Bu,L., De Leon,D.D. and De Leon,M. (2008) Expression of E-FABP in PC12 cells increases neurite extension during differentiation: involvement of n-3 and n-6 fatty acids. *J. Neurochem.*, **106**, 2015–2029.
58. Robson,L.G., Dyall,S., Sidloff,D. and Michael-Titus,A.T. (2010) Omega-3 polyunsaturated fatty acids increase the neurite outgrowth of rat sensory neurones throughout development and in aged animals. *Neurobiol. Aging.*, **31**, 678–687.
59. De Leon,M., Welcher,A.A., Nahin,R.H., Liu,Y., Ruda,M.A., Shooter,E.M. and Molina,C.A. (1996) Fatty acid binding protein is induced in neurons of the dorsal root ganglia after peripheral nerve injury. *J. Neurosci. Res.*, **44**, 283–292.
60. Alon,U. (2007) Network motifs: theory and experimental approaches. *Nat. Rev. Genet.*, **8**, 450–461.
61. Shen-Orr,S.S., Milo,R., Mangan,S. and Alon,U. (2002) Network motifs in the transcriptional regulation network of Escherichia coli. *Nat. Genet.*, **31**, 64–68.
62. Robertson,G., Bilenky,M., Lin,K., He,A., Yuen,W., Dagpinar,M., Varhol,R., Teague,K., Griffith,O.L., Zhang,X. *et al.* (2006) cisRED: a database system for genome-scale computational discovery of regulatory elements. *Nucleic Acids Res.*, **34**, D68–D73.